1st International Conference on Experiments/Process/System Modeling/Simulation/Optimization
1st IC-EpsMsO
Athens, 6-9 July, 2005

NETWORK TRAFFIC MODELING AND PREDICTION USING MULTIPLICATIVE SEASONAL ARIMA MODELS

Vassilios C. Moussas[1*], **Marios Daglis**[1], **and Eva Kolega**[1]

[1] Network Operations Centre (NOC), Technological Educational Institution (T.E.I.) of Athens
Egaleo GR-12210, Greece, [*]e-mail: vmouss@teiath.gr

Keywords: Time-Series, ARIMA, Seasonal, Network, Traffic, Utilization.

Abstract: *Today's network designers are expected to plan for future expansion and to estimate the network's future utilization. Several simulators can be used for 'what-if' scenarios but they all require as input some estimates of the future network use. A method for estimating the future utilization of a network is presented in this work. Network utilization is initially modeled using an ARIMA model (p, d, q), but its prediction accuracy has a limited time span. The prediction is improved significantly by using a multiplicative seasonal ARIMA (p, d, q) x $(P, D, Q)_s$ model. The seasonal model proved extremely capable to recreate the current data and predict the future utilization with precision. The only requirement of the nonlinear model is the availability of longer past records. The daily, weekly and monthly datasets were collected from real-life network utilization, at the TEI of Athens campus network.*

1 INTRODUCTION

Network traffic monitoring and future traffic predictions play a significant role in the design of networks and network upgrades. It is important to forecast the network traffic workload when planning, designing, controlling and managing various LAN, MAN or WAN networks. Time-series techniques look promising for modeling network utilization, as, the Auto Regressive Integrated Moving Average models are able to capture trends and periodic behavior.

The ARIMA and Seasonal ARIMA models have been applied successfully in economy, signal processing, safety and other research fields [1, 2]. Recently, several researchers modeled specific network characteristics, ranging from long-term backbone traffic forecasting to wireless and GSM traffic, using time series models with success [3, 4, 5].

In this paper a set of ARIMA models is used to predict future network utilization from simple and common traffic monitoring data. The aim is to obtain accurate predictions using common (easy to find, simple to apply and widely used) traffic datasets. Four different "utilization" cases were studied, namely, a campus Internet connection, a building backbone, an office LAN, and a 160 lines dialup connection.

2 TRAFFIC DATA COLLECTION

There are several types of data to collect, when studying the traffic of a network. Almost any traffic characteristic may be measured and logged i.e. bit or packet rate, packet size, protocols, ports, connections, addresses, server load, applications, etc. Routers, firewalls, servers and managers (servers with agents) can be used for this task, but measuring and archiving all these data for potential future treatment is not a regular procedure. Usually most networks log only the load of their lines and the utilization some critical resources, using a more detailed monitoring only when a resource requires specific attention. As a result, traffic rate or utilization are the only data collections that are always available and with long history records on almost any network.

In this work we focus on such commonly found data sets in order to develop a more general and widely applied tool. Our traffic data are taken from the router's standard MIB or from the server's typical. In order to avoid device or system specific problems the data were taken via a monitoring tool. The Multi Router Traffic Grapher

(MRTG®) [8] tool is used for all data collection. The MRTG tool is very common, widely applied, and easily implemented software for collecting and monitoring utilization data from a router or server MIB. The MRTG tool produces standard log files with current and past data that can be downloaded and saved by any browser or simple GET commands. Our method uses some MATLAB® procedures that read those standard files and prepare the data for the model identification steps.

We selected 4 different types of load/utilization data regularly collected in our campus network (Figure 1): a) the TEI of Athens campus Internet connection traffic rate (in bps), b) one of the campus buildings backbone traffic rate (in bps), c) an offices-only VLAN with no labs (in bps), d) the number of the on-line Dialup users.

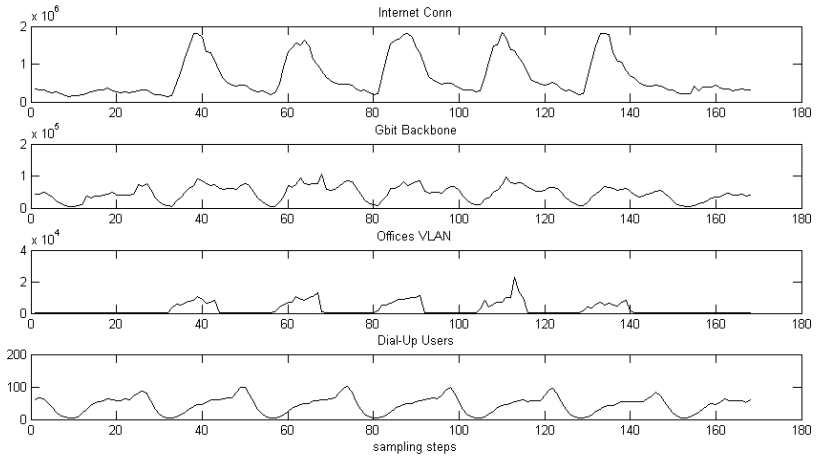


Figure 1. Average utilization data from the TEI of Athens campus network (weekly data). a) The campus Internet connection. b) A building backbone. c) An offices only VLAN. d) On line dial-up users.

Data sampling is performed by default every 5 minutes. Although it can be changed, this is a standard adjustment for MRTG and it is used in most implementations. Daily traffic is therefore expressed in steps of a 5-minute average traffic rate, measured in bits per seconds (bps). Sampling by a shorter period (e.g. 1 min) may sometimes lead to inaccuracies due to timing and delayed responses from the router's MIB and sampling by a longer period (e.g. 10 min) may lead to loss of information due to averaging. Therefore, the default sampling rate (5-min) was finally used for this work.

For visualizing long periods MRTG does some averaging over a longer step. Weekly traffic is expressed in 30-min steps, monthly traffic in 2-hour steps, and, yearly traffic in 1-day steps. In order to preserve some information lost by averaging the tool logs together the average and the maximum value for each step. They can both be used to produce forecasts, regarding the average or the maximum behavior respectively.

As it also clear in Figure 1, the utilization of the network resources demonstrates a daily periodicity. In addition most traffic patterns demonstrate another, less apparent, weekly periodicity. The monthly graph in Figure 2 clearly presents both types of periodicity. Such characteristics are important for the modeling procedure that follows, and for the forecasting accuracy of the method.

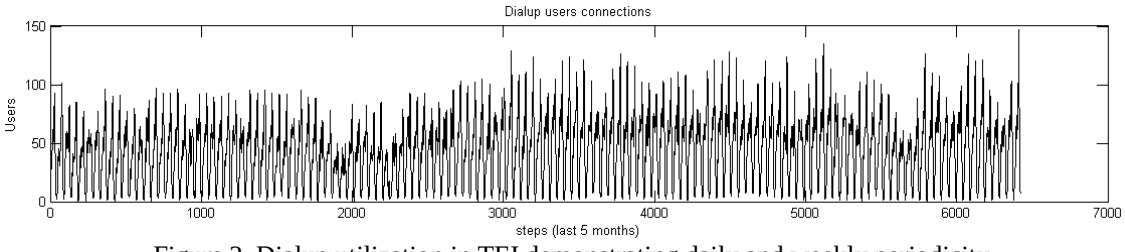


Figure 2. Dialup utilization in TEI demonstrating daily and weekly periodicity.

3 TIME SERIES MODELING

The principle underlying this methodology is that traffic data occur in a form of a time series where observations are dependent. This dependency is not necessarily limited to one step (Markov assumption) but it can extend to

many steps in the past of the series. Thus in general the current value X_t (= network traffic at time t) of the process X can be expressed as a finite linear aggregate of previous values of the process and the present and previous values of a random input u [1], i.e.

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \cdots - \theta_q u_{t-q} \quad (1)$$

In this equation … and … represent respectively the traffic volume and the random input at equally spaced time intervals t, t-1, t-2, … .The random input u constitute a white noise stochastic process, whose distribution is assumed to be Gaussian with zero mean and standard deviation σ_u.

Eqn. (1) can be economically rewritten as (4) by defining the autoregressive (AR) operator of order p and the moving-average (MA) operator of order q respectively by (2) & (3):

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \quad (2)$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p \quad (3)$$

$$\varphi(B) X_t = \theta(B) u_t \quad (4)$$

where, B stands for the backward shift operator defined as $B^s X_t = X_{t-s}$. Another closely related operator is the backward difference operator ∇ defined as $\nabla X_t = X_t - X_{t-1}$ and thus, $\nabla = 1 - B$, $\nabla^d = (1-B)^d$ and $\nabla_s^D = (1-B^s)^D$.

The autoregressive moving-average model (ARMA) as formulated above is limited to modeling phenomena exhibiting stationarity. Clearly this is not the case for the network traffic data of Fig. 1. It is possible though that the processes still possess homogeneity of some kind. It is usually the case that the d^{th} difference of the original time series exhibits stationary characteristics. The previous ARMA model could then be applied to the new stationary process ∇X and eqn. (4) will correspondingly read

$$\varphi(B) \nabla^d X_t = \theta(B) u_t \quad (5)$$

This equation represents the general model used in this work. Clearly, it can describe stationary ($d = 0$) or non-stationary ($d \neq 0$), purely autoregressive ($q = 0$) or purely moving-average ($p = 0$) processes. It is called autoregressive integrated moving-average (ARIMA) process of order (p, d, q). It employs p + q +1 unknown parameters $\varphi_1, \ldots, \varphi_p$; $\theta_1, \ldots, \theta_q$; σ_u , which will have to be estimated from the data.

Starting from ARIMA model of eqn. (5) it can be deducted [1] that a seasonal series can be mathematically represented by the general multiplicative model often called Seasonal ARMA or SARIMA

$$\varphi_p(B) \Phi_P(B^s) \nabla^d \nabla_s^D X_t = \theta_q(B) \Theta_Q(B^s) u_t \quad (6)$$

The general scheme for determining a model includes three phases, which are:

1. Model identification, where the values of the parameters p, d, q are defined
2. Parameter estimation, where the $\{\varphi\}$ and $\{\theta\}$ parameters are determined in some optimal way, and
3. Diagnostic checking for controlling the model's performance.

As is stated however in [1], there is no uniqueness in the ARIMA models for a particular physical problem. In the selection procedure, among potentially good candidates one is aided by certain criteria. Although more advanced methods for model selection have been proposed [6] the most common and classic criteria remain the Akaike's Information Criterion (AIC) and the Schwartz's Bayesian Information Criterion (SBC or BIC) [1,7]. If $L = L(\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q, \sigma_u)$ represents the likelihood function, formed during the parameter estimation, the AIC and SBC are expressed respectively, as

$$\text{AIC} = -2 \ln L + 2k \quad \& \quad \text{SBC} = -2 \ln L + \ln(n) k \quad (7)$$

where, k = number of free parameters (= p + q) and n = number of residuals that can be computed for the time series. Proper choice of p and q calls for a minimization of the AIC and SBC.

4 ARIMA MODEL BUILDING

In the search of the parameters *p*, *d*, *q* of the ARIMA model of eqn. (5) we work first with a single day traffic data from Fig. 2. The curve in fig. 2 represents the average daily traffic of the TEI campus Internet connection. The second difference of the daily curve (i.e. $d = 2$) demonstrates stationarity, as it is shown by its Autocorrelation Function (ACF). The original curve, its 1st and 2nd differences and their corresponding ACF are shown in Fig. 3.

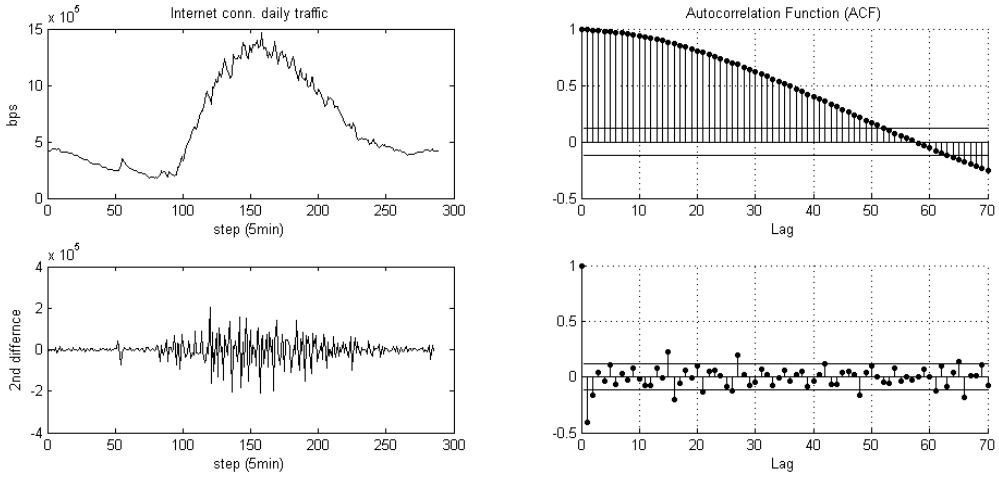


Figure 3. Internet connection of the TEI campus. Daily traffic data, the 2nd difference and their ACFs.

The criteria are minimized for $p = q = 3$, therefore the ARIMA (3,2,3) model should be used. Parameter estimation for this model yields the following values: $\varphi_1 = -0.317$, $\varphi_1 = 0.318$, $\varphi_1 = -0.153$, $\theta_1 = 0.764$, $\theta_1 = 0.899$, and, $\theta_1 = -0.7475$. The model fits satisfactorily the data, but it may predict accurately only for a few steps in the future (Figure 4). As shown in Figure's (4) inside plot, it is not reliable for longer predictions.

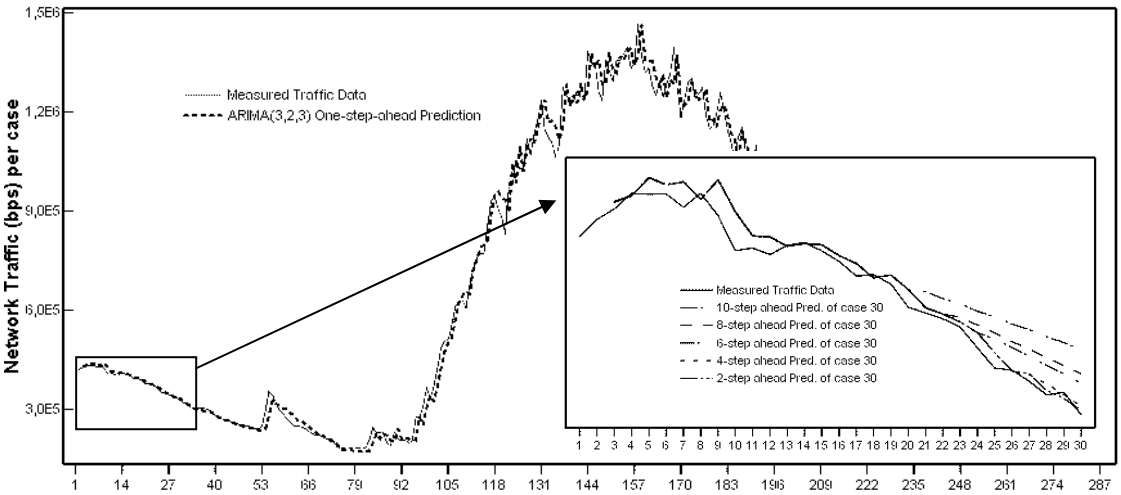


Figure 4. Fitting of the average daily traffic (Internet connection) using the ARIMA (3,2,3) model, and, its predictions. Short time (1-4 steps) forecasts of the 30th case may be acceptable, but 30 min (6 steps) or longer forecasts for the same case are inaccurate.

Data set	(*p*,*d*,*q*)	AR & MA Parameters
Internet Conn.	(3,2,3)	AR: -0.317, 0.318, -0.153, MA: 0.764, 0.899, -0.7475
Gbit Backbone	(2,2,1)	AR: -0.4334, -0.2827, MA: 0.9676
Offices VLAN	(1,1,1)	AR: -0.3864, MA: -0.5468
Dial-Up Users	(1,2,1)	AR: -0.1833, MA: 0.5915

Table 1 : ARIMA model parameters for all 4 traffic datasets

The identified models that minimize the AIC/SBC criteria for each dataset are shown in Table 1. The same conclusions regarding future predictions hold for all 4 kinds of utilization data. The ARIMA model is not accurate for long forecast periods. As already mentioned in section 2 the utilization data demonstrate some periodicities and therefore another type of model must be applied that will take into account such properties.

5 SEASONAL ARIMA MODEL BUILDING

As the above approach yields acceptable predictions only for a short period and in order to predict reliably for longer periods, the use of the seasonal approach is required. From the first ACF plot in figure 5 becomes clear that the data set has two seasonal components, a daily one (at 24h) and a weekly one (at 168h).

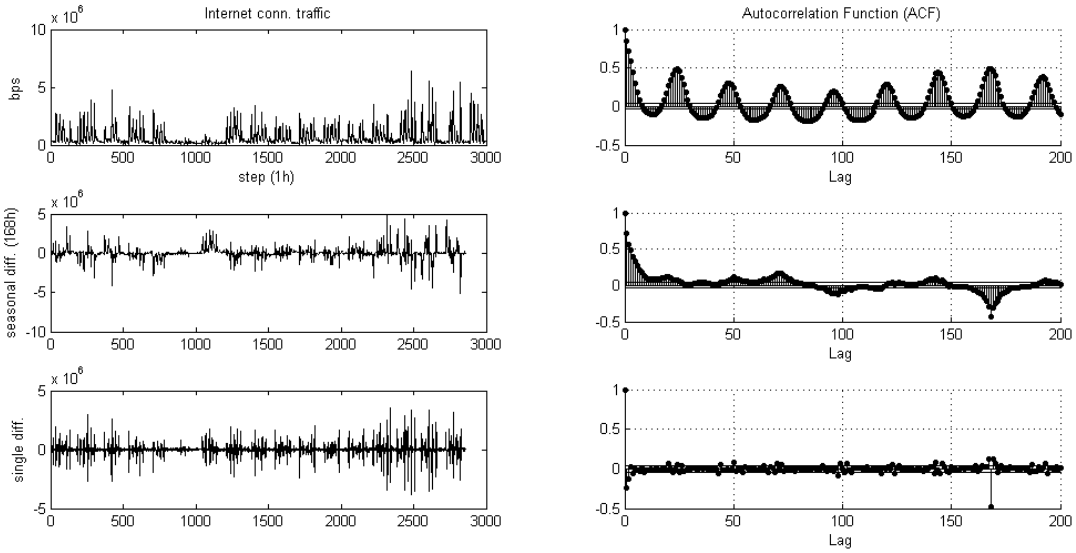


Figure 5. Internet connection. Four months of traffic data, single & seasonal differences and their ACFs.

In the search of the parameters *p, d, q* and *P, D, Q* of the SARIMA model of eqn. (6) we work as before by taking one single and one seasonal difference. After two differences (i.e. $D = 1$ & $d = 1$) the ACF demonstrates stationarity and presents a peak at the seasonal component, i.e. lag 168, when the weekly periodicity is selected (Fig. 5).

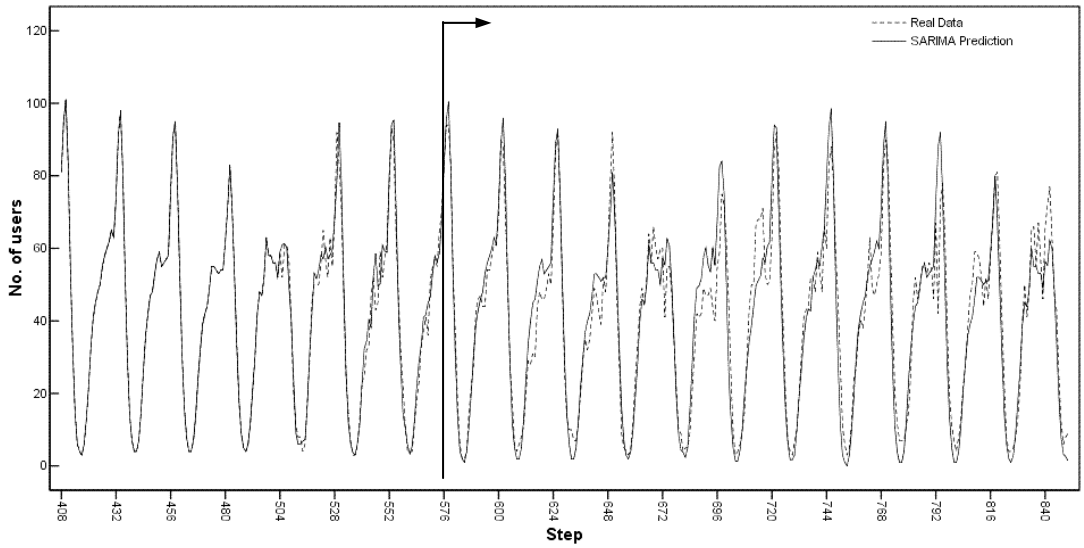
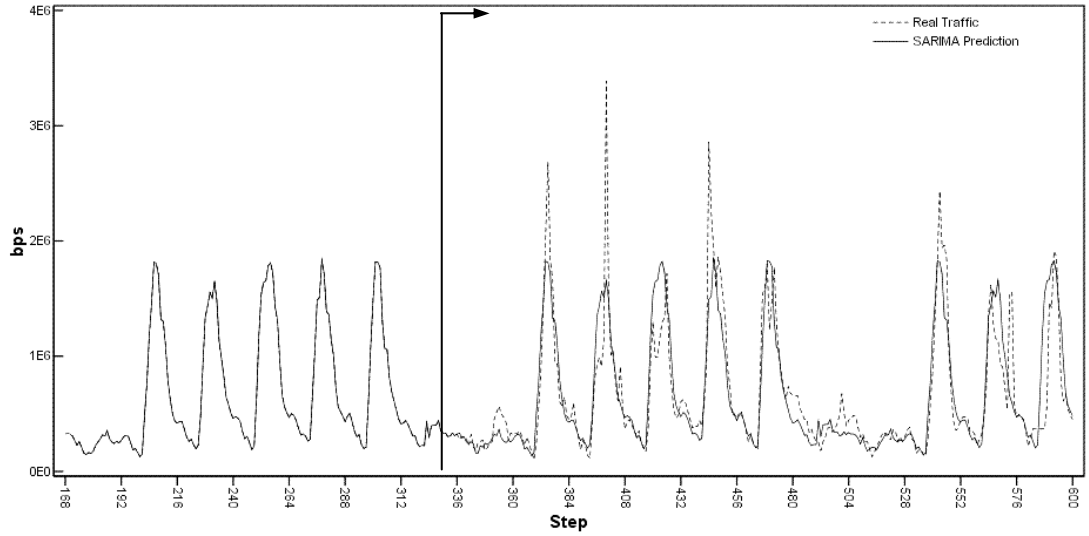
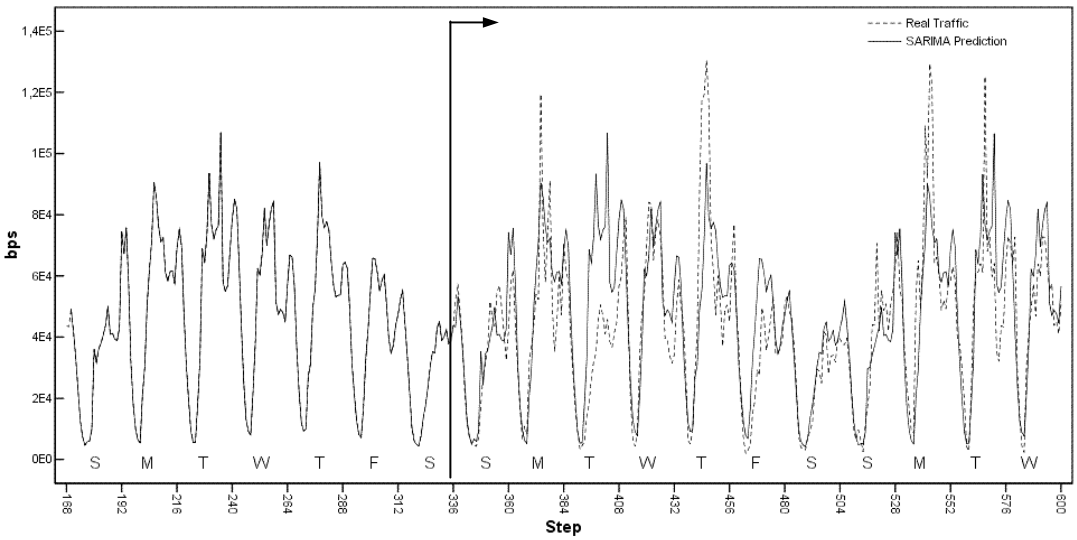


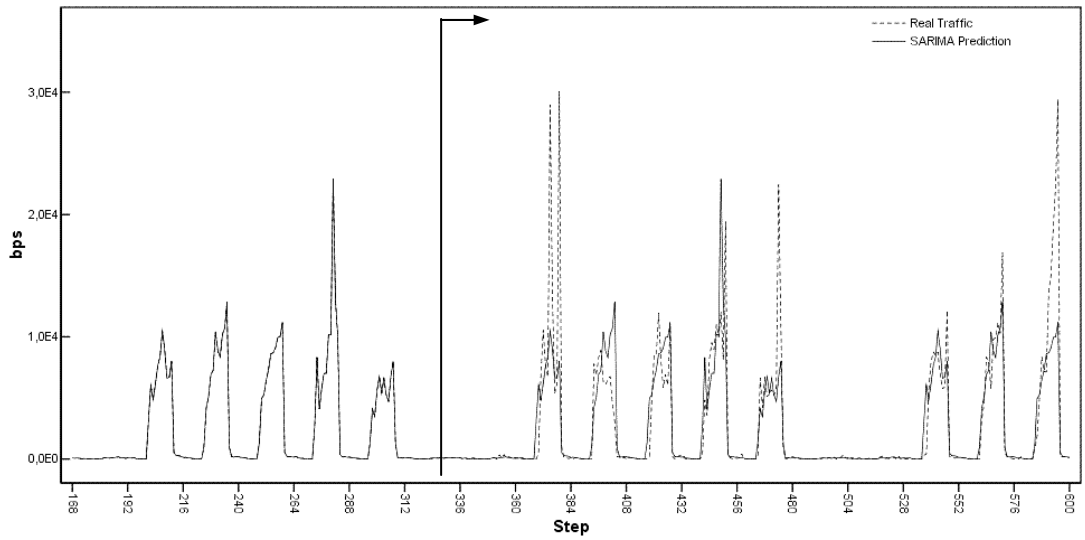
Figure 6. Prediction of the weekly Dial-Up utilization using the SARIMA $(1,1,1)\times(0,1,2)_{168}$ model. Starting from step 577 the model predicts satisfactorily in the future.



(a) Campus Internet Connection



(b) Building Backbone



(c) Offices only VLAN

Figure 7. Predictions of the weekly traffic using the SARIMA $(1,1,1)x(0,1,1)_{168}$ model. Starting from step 350 the model predicts satisfactorily for 10 days in the future, including days with lower traffic.

Table 2 summarizes the results for each dataset using seasonal models with weekly periodicity. Each model is also used for future predictions and the results are quite accurate as shown in figures 6 & 7.

Data set	$(p,d,q)x(P,D,Q)_s$	AR, MA & Seasonal MA Parameters
Internet Conn.	$(1,1,1)x(0,1,1)_{168}$	AR: 0.520607, MA: 0.975213, SMA: 0.767889
Gbit Backbone	$(1,1,1)x(0,1,1)_{168}$	AR: 0.404779, MA: 0.982133, SMA: 0.921960
Offices VLAN	$(1,1,1)x(0,1,1)_{168}$	AR: 0.485866, MA: 0.996179, SMA: 0.921947
Dial-Up Users	$(1,1,1)x(0,1,2)_{168}$	AR: 0.609797, MA: 0.988091, SMA: 0.81489, 0.09942

Table 2 : Seasonal ARIMA model parameters for all 4 traffic datasets

In addition to the weekly models, SARIMA models with daily periodicity were also tested. They predict equally well the future daily utilization, but under the assumption that past and future day belong in the same category. As it is clear in figure 6, there are two types of daily utilization the one of a normal working day and the other of a weekend or vacation. Daily SARIMA doesn't separate and performs an average prediction for the next day whatever the day is. On the other hand the weekly SARIMA handles the traffic pattern of an entire week that also includes the lower utilization observed during weekends. This is most obvious in figure 7c showing the "Offices VLAN" traffic when the offices are closed on Saturday and Sunday.

For the above reasons we selected the weekly seasonal ARIMA for weekly, monthly and long time predictions. The daily seasonal ARIMA is more appropriate for a different type of traffic monitoring, i.e., by having several daily patterns available, one can monitor traffic and adaptively select the current type of utilization. Of course this is not restricted to working days or weekends, but also to other types of traffic such as connection problems, anomalies or intrusions. Such an online identification of the traffic pattern is beyond the scope of this work (modeling & prediction) and it will be studied in a separate contribution.

6 CONCLUSIONS

In this work we studied the fitting and forecasting capabilities of time series models to the real traffic and utilization data of a campus network. The multiplicative seasonal ARIMA model produced satisfactory forecasts for four different types of utilization. All future predictions were accurate once the network is working under normal conditions (no failures, DoS attacks, etc.). Moreover the seasonal model with weekly periodicity was superior to the one with daily periodicity, as it takes into account the traffic reduction during weekends. Our results agree to those of other researchers and we concluded that the so-called seasonal time series could offer a unique tool for modeling, with particular bearing to the real time monitoring and prediction of network traffic & utilization. A further exploration of the above models for online identification follows in a separate contribution.

REFERENCES

[1] Box G., Jenkins G.M. and Reinsel G. (1994), *Time Series Analysis: Forecasting & Control*, 3rd ed. Prentice Hall,.

[2] Solomos G.P. and Moussas V.C. (1991), "A Time Series Approach to Fatigue Crack Propagation", Structural Safety, 9, pp.211-226.

[3] Papagiannaki K., Taft N., Zhang Z. and Diot C. (2003), "Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models", *IEEE Infocom* 2003.

[4] Shu Y., Yu M., Liu J. and Yang O.W.W. (2003), "Wireless Traffic Modeling and Prediction Using Seasonal ARIMA Models", *IEEE Intl. Conference on Communication May 2003*, ICC'03 vol. 3.

[5] You C. and Chandra K. (1999), "Time Series Models for Internet Data Traffic", *In Proc. 24th Conf. on Local Computer Networks October 1999*, LCN-99.

[6] Katsikas S.K., Likothanassis S.D. and Lainiotis D.G. (1990), "AR model identification with unknown process order", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-38, No. 5, pp. 872-876.

[7] Akaike H. (1969), "Fitting Autoregressive models for Prediction", Ann. Inst. Sat. Math., vol. 21, pp. 243-247.

[8] Oetiker Tobias, (2005), Multi Router Traffic Grapher (MRTG) tool, Software Package & Manuals, http://people.ee.ethz.ch/~oetiker/webtools/mrtg/, Last visit: Feb 2005.